

Enhancing Linguistic Competence of Language Models through Pre-training with Language Learning Tasks

Atsuki Yamaguchi Maggie Mi Nikolaos Aletras

School of Computer Science, University of Sheffield, United Kingdom

{ayamaguchi1, zmi1, n.aletras}@sheffield.ac.uk

Abstract

Language models (LMs) are pre-trained on raw text datasets to generate text sequences token-by-token. While this approach facilitates the learning of world knowledge and reasoning, it does not explicitly optimize for linguistic competence. To bridge this gap, we propose **L2T**, a pre-training framework integrating **Language Learning Tasks** alongside standard next-token prediction. Inspired by human language acquisition, L2T transforms raw text into structured input-output pairs to provide explicit linguistic stimulation. Pre-training LMs on a mixture of raw text and L2T data not only improves overall performance on linguistic competence benchmarks but accelerates its acquisition, while maintaining competitive performance on general reasoning tasks.¹

1 Introduction

Language models (LMs) are pre-trained using causal language modeling (CLM), a next-token prediction objective (Radford et al., 2018). While CLM pre-training equips LMs with world knowledge and general capabilities (Brown et al., 2020; Kojima et al., 2022), it does not optimize for linguistic competence, i.e., the capacity to comprehend and interpret diverse linguistic phenomena (Chomsky, 1965; Waldis et al., 2024).

Consequently, LMs often behave as stochastic parrots (Bender et al., 2021). They mimic surface-level patterns without grasping the underlying linguistic scaffolding (Chang and Bergen, 2024; López-Otal et al., 2025). This phenomenon mirrors rote learning in humans (Plunkett and Marchman, 1993), where learners reproduce patterns without understanding the generative rules (Brown and Berko, 1960; Myles et al., 1998).

We hypothesize that pre-training on **language learning tasks**, which require processing beyond

¹Our code and models are available via <https://github.com/gucci-j/l2t>.

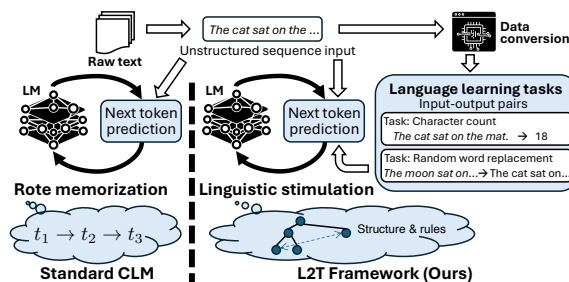


Figure 1: L2T vs. standard CLM over raw text.

raw sequence reconstruction, can stimulate the development of linguistic competence. Specifically, we target data-efficient acquisition of morphological, syntactic, and semantic knowledge. Inspired by human language acquisition, this approach encourages the development of structured representations that go beyond surface-level co-occurrence (Alishahi, 2011; Perfors et al., 2011; Culbertson and Adger, 2014). Evidence suggests that both humans and neural networks benefit from structured linguistic input during learning (Elman, 1993; Galke et al., 2024; Hu et al., 2025).²

We propose **L2T**, a pre-training framework integrating **Language Learning Tasks** alongside standard CLM (Figure 1). Unlike instruction tuning requiring external supervision, L2T induces structure directly from raw text. By converting text into structured input-output pairs, it enables the model to learn dependencies and how to restructure information, providing explicit linguistic stimulation absent in standard CLM. We evaluate L2T by pre-training LMs on a mixture of L2T data and raw text at scales of 500M and 1B parameters. Our main **contributions** are: (1) the L2T framework substantially improves linguistic competence up to 11.3% (2.8% on average) on BLiMP (Warstadt et al., 2020), while retaining general performance; and (2) empirical evidence that L2T accelerates this process.

²Appendix A provides a detailed review of related work.

Type	Tasks and Examples	Links
Char	Char Count: count characters (Text \rightarrow 4)	§B.1.1
	Masked Char: reconstruct masked characters (c_ar \rightarrow char)	§B.1.2
	Space: restore whitespace (Ilikea \rightarrow I like a)	§B.1.3
	Typo: correct synthetic typos (typ0 \rightarrow typo)	§B.1.4
Word	Last: predict concluding phrase ([Text] _{A/B} \rightarrow A)	§B.2.1
	Masked Word: reconstruct tokens (I [MASK] \rightarrow I am)	§B.2.2
	Random: correct tokens (Sea am \rightarrow I am)	§B.2.3
	Shuffle: restore word order ($w_1 w_3 w_2 \rightarrow w_1 w_2 w_3$)	§B.2.4
	Token Type: count linguistic categories ([Text] _{Digit} $\rightarrow n \in \mathbb{Z}$)	§B.2.5
Sent	Deletion: remove unrelated sentence (A [X] C \rightarrow AC)	§B.3.1
	Reordering: restore sentence flow ($S_3 S_1 S_2 \rightarrow S_1 S_2 S_3$)	§B.3.2
Disc	Fill Middle: fill-in-the-middle for passages ($P_1 ? P_3 \rightarrow P_2$)	§B.4.1
	Half: complete the latter half ([Start]... \rightarrow [End])	§B.4.2
	One: generate from one-word prefix ([Word] \rightarrow [Text])	§B.4.3

Table 1: The 14 L2T tasks with links to the detailed task definitions and examples in Appendix B.

2 L2T: Language Learning Tasks

The L2T framework comprises 14 language learning tasks (Table 1) designed to provide explicit structural stimulation. Each task converts a raw text segment into a structured input-output pair (x, y) , where x is a linguistically perturbed or queried input, and y is the restored or analyzed output.

While prior work has often relied on architectural modifications (Xu et al., 2021) or complex curricula (Hu et al., 2025; Oba et al., 2023; Mi, 2023), our aim is distinct: *we investigate how structured stimulation impacts pre-training dynamics and the development of linguistic competence*. We hypothesize that the *rote learning* of LMs stems in part from the single-task nature of CLM, which prioritizes surface-level statistics over structural understanding. In contrast, humans do not acquire language by optimizing a single objective; they learn through multiple tasks (Spelke, 2022).

To mirror this multi-task learning, L2T generates various tasks by leveraging raw text across four levels of linguistic granularity, providing supervision signals without external resources. Prior work (Carroll et al., 1992) shows that error correction in human learners improves morphological awareness, particularly for previously encountered forms. Accordingly, we use character-level (Char) tasks (e.g., *Typo*) to target subword features, discouraging surface-level matching while enhancing morphological awareness. Word- and sentence-level tasks (e.g., *Shuffle*) disrupt linear order, promoting structural inference over sequential statistics, consistent with word-reordering studies (Akhtar, 1999; Chomsky, 2002; Zhang and Clark, 2015). Finally, discourse-level (Disc) tasks (e.g., *Fill Middle*) require completion across longer context, supporting global coherence and ambiguity resolution. Similar completion tasks benefit human language learn-

ers (Keating, 2008). By integrating these diverse signals, L2T establishes the structural scaffolding required for linguistic competence, complementing world knowledge acquired through standard CLM.

3 Experimental Setup

3.1 Pre-training Data Scenarios

We derive our pre-training corpora from the English FineWeb-Edu dataset (Penedo et al., 2024). To evaluate the L2T framework under different resource constraints, we establish two distinct data scenarios (**Disjoint** and **Shared**) to determine if L2T yields consistent benefits across varying volumes of unique source text. Following Cheng et al. (2024a), we fix the total training budget at 100B tokens. Crucially, this budget exceeds the compute-optimal thresholds defined by the Chinchilla scaling laws (Hoffmann et al., 2022), enabling us to evaluate performance in a regime characterized by sufficient token availability relative to model size.

Disjoint (Abundant Data). This configuration simulates a regime where high-quality source text is plentiful. We partition an initial 100B token sample into two mutually exclusive subsets of equal size. The first subset is retained for standard CLM, while the second is used for generating L2T samples. As the combination of raw and L2T samples exceeds the budget, we subsample the mixture to adhere to the 100B token limit. The resulting Disjoint dataset comprises approximately 36B raw and 64B L2T tokens. This setup evaluates the impact of the framework when the model has access to high document diversity alongside structured tasks.

Shared (Limited Data). This configuration targets settings where source text availability is constrained. We use a fixed set of 42B source tokens both as raw text for CLM and L2T sample generation. This combination results in a total of 100B tokens.³ This strategy decouples the impact of task structure from data diversity, testing if L2T can improve linguistic inductive bias without the introduction of unique new documents.

3.2 Training Configuration

We pre-train from scratch 500M and 1B Qwen2.5-based models (Qwen Team et al., 2025). These models use the Mistral (Jiang et al., 2023) tokenizer with a 32K English-centric vocabulary. This

³While multiple transformations are possible, we assign a single, random L2T task to each sample for brevity.

	Data	Semantics		Morphology				Syntax					Overall		
		Quant	NPI	Ana Agr	Irregul	DN Agr	SV Agr	Arg Str	Bind	Ctrl Rais	Ellips	Fill Gap		Island	
500M	Disj.	Raw	65.6	76.5	93.8	91.8	93.1	86.0	78.2	72.8	78.5	87.1	77.5	63.0	78.6
		L2T	71.4	78.3	96.0	93.2	92.4	88.6	78.6	75.5	79.5	86.0	75.4	70.8	80.2
	Shar.	Raw	67.5	68.8	94.0	92.3	93.6	86.5	78.6	77.9	79.6	88.5	75.6	60.6	78.1
		L2T	74.9	78.7	97.9	96.2	93.4	88.4	80.5	74.2	81.4	86.2	76.4	68.1	80.9
1B	Disj.	Raw	71.1	78.6	95.3	89.5	94.5	84.8	78.0	78.6	79.3	86.7	75.4	60.2	79.0
		L2T	75.2	77.9	96.2	92.7	93.3	88.3	79.4	76.1	80.8	86.5	74.9	71.5	80.8
	Shar.	Raw	74.8	71.2	95.7	90.8	94.6	87.4	78.2	75.7	78.5	86.5	77.0	61.7	78.9
		L2T	72.1	82.0	97.0	94.2	92.7	88.4	80.0	76.7	82.0	84.5	77.0	68.6	81.2

Table 2: Linguistic competence on BLiMP. Green highlights better performance of L2T (ours) over Raw.

replaces the 130K vocabulary of Qwen2.5 to reduce compute cost. We set the sequence length to 2,048 and batch size to 256. To ensure a fair comparison, we compute loss on all tokens.

3.3 Evaluation Framework

Baselines. We pre-train models (Raw; 500M and 1B) on 100B and 42B raw tokens for the Disjoint and Shared settings, respectively. For fair comparison, we use a budget of 100B tokens in all settings. Hence, the Shared Raw model trains over the same 42B tokens multiple times to meet this quota.⁴

Linguistic Competence. We use BLiMP to measure linguistic competence following Shah et al. (2024). This benchmark covers 12 linguistic phenomena across 67 datasets within: Semantics, Morphology, and Syntax. Appendix C.3 provides details and their associated linguistic constraints. We prioritize zero-shot log-likelihood comparisons of minimal pairs to avoid prompting and probing biases (Hu and Levy, 2023; Alajrami and Aletras, 2022; Belinkov, 2022; Levy et al., 2023).

General Benchmarks. We use: Reading Comprehension (RC; RACE, SciQ, LogiQA); Commonsense Reasoning (CR; ARC-Easy, COPA, OpenBookQA, SIQA, HellaSwag); and Language Modeling (LAMBADA). We also include ReCoRD, which combines RC and CR.⁵ Since the volume of raw text remains constant across settings, this aims to confirm whether L2T tasks are complementary or being detrimental to CLM on raw text.

4 Results

4.1 Linguistic Competence

Table 2 shows the performance on BLiMP for all models. L2T consistently enhances linguistic

⁴Baselines and L2T models use the same hyperparameters (Appendix C.2).

⁵Evaluation metrics include normalized accuracy for ARC, HellaSwag, LogiQA, OBQA, and SciQ; F1 for ReCoRD; and standard accuracy for the remaining tasks. We apply five-shot prompting for ARC, LogiQA, OBQA, and SIQA, while using zero-shot for the rest. We use a single deterministic run.

competence. In the Disjoint setup, L2T models outperform the Raw baselines at both 500M and 1B scales, achieving overall scores of 80.2 (500M) and 80.8 (1B), compared to Raw baseline scores of 78.6 and 79.0. Notably, this performance gap widens in the Shared scenario, where L2T models surpass the Raw baselines by 2.8 and 2.3 points. This reinforces our human learning analogy (§2): while Shared Raw mimics “rote learning,” L2T mirrors multi-task acquisition by applying diverse structural lenses to the same data. Our results confirm that linguistic competence depends not just on data volume, but on the diversity of signals applied during training.

We further analyze the effects across different linguistic subfields and phenomena. L2T consistently enhances linguistic competence across model sizes and data scenarios, improving six phenomena across all subfields. Island effects (Island) exhibit the most substantial gains, ranging from 6.9 (1B; Shared) to 11.3 points (1B; Disjoint). Detailed analysis (§5 and Appendix E) reveals that nearly all L2T tasks contribute to this gain. This suggests that structured tasks across varying granularity, rather than solely local transformations, provide a structural scaffolding for capturing long-distance dependencies.

Conversely, L2T offers no improvement for determiner-noun agreement (DN Agr) or ellipsis (Ellips). These results serve as a diagnostic of our method: performance for these phenomena appears to reach saturation, as Raw models achieve high accuracy (93+ and 86+, respectively) through implicit learning of frequent patterns (Tajeddin and Rahimi, 2017; Khullar et al., 2020). This result aligns with the finding of Shah et al. (2024), who report that LMs are substantially accurate on morphological tasks followed by syntactic and semantic tasks. However, limits to this scaffolding remain. Complex phenomena, such as Filler-Gap dependencies (Fill Gap), show no improvement from individual tasks (see Appendix E), indicating that capturing such structures is more challenging and

	Data	Reading Comprehension			RC+CR	Commonsense Reasoning					Language Modeling		
		RACE	SciQ	LogiQA	ReCoRD	ARC	COPA	OBQA	PIQA	SIQA	HellaSwag	LAMBADA	
500M	Disj.	Raw	30.0	67.9	23.8	62.9	57.4	66.0	31.4	63.9	38.1	37.8	30.6
		L2T	29.8	67.3	26.6	60.7	56.4	64.0	30.8	62.9	38.3	35.2	28.2
	Shar.	Raw	28.3	66.1	25.0	62.2	56.6	66.0	29.0	64.0	38.1	37.1	30.6
		L2T	29.5	63.0	26.9	62.0	57.7	64.0	29.6	65.1	38.5	36.2	30.2
1B	Disj.	Raw	29.6	72.4	24.4	64.9	60.4	61.0	31.0	65.9	38.6	39.7	32.7
		L2T	29.8	70.2	27.2	63.1	58.4	66.0	32.0	65.9	38.5	37.8	30.9
	Shar.	Raw	29.5	68.0	26.1	65.3	60.6	69.0	31.2	66.3	38.4	39.4	31.9
		L2T	30.2	68.0	26.7	63.6	56.4	65.0	30.0	64.3	37.4	37.6	31.3
Random guessing		25.0	25.0	25.0	19.1	25.0	50.0	25.0	50.0	33.3	25.0	0.0	

Table 3: General benchmark performance. Green denotes better performance of L2T (ours) over Raw.

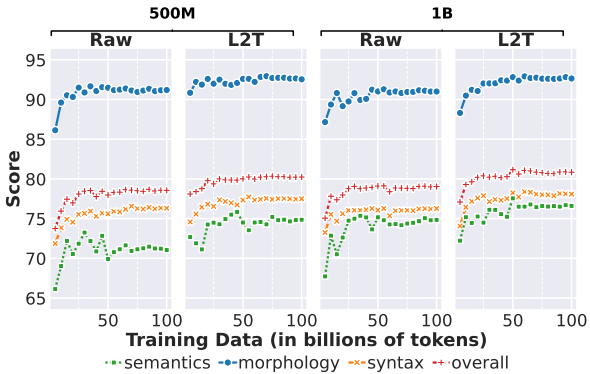


Figure 2: Accuracy by linguistic subfield in BLiMP between Raw and L2T across model sizes and training steps using Disjoint Raw and L2T data.

may require targeted discourse-level objectives.

Finally, Figure 2 illustrates the development of linguistic competence throughout the pre-training process. Across subfields, substantial improvements occur during the initial 20-30B tokens of training, with gains continuing and leveling off around 50B tokens. These results reveal similar trajectories as those identified by Shah et al. (2024) for Pythia models trained on raw text. However, L2T models consistently outperform Raw models across all linguistic subfields from the earliest stages of training, evident as early as 5B tokens. This indicates that L2T effectively boosts learning in “the window of maximal development”, the period where the model improves its cognitive abilities linearly (Shah et al., 2024). At this 5B token mark, L2T models show an average performance advantage over Raw models ranging from 3.3 (Syntax) to 6.5 (Semantics) for the 500M scale, and from 1.0 (Syntax) to 4.5 (Semantics) for the 1B scale. This faster acquisition indicates that L2T tasks function as a force multiplier for efficiency by introducing linguistic inductive biases complementary to CLM. By effectively stimulating language learning, L2T establishes a durable advantage, preserving a performance gap that the implicit

acquisition of standard Raw models cannot close.

4.2 General Benchmarks

Table 3 compares the performance of L2T and Raw models on general benchmarks.

Disjoint. In the Disjoint setup, L2T models maintain competitive performance relative to Raw. The 500M model shows an average drop of 0.87 points, while the 1B model shows a negligible average difference of 0.07 points. This confirms that general reasoning performance remains stable when the model retains access to a large volume of unique documents.

Shared. In the Shared configuration, the impact of L2T varies by model scale. While the 500M model achieves an average improvement of 0.15 points, the 1B model experiences a performance drop of 1.38 points, primarily in CR tasks such as ARC (-4.2). This divergence highlights a tension between linguistic structure learning and factual reinforcement (Fedorenko and Varley, 2016). The Shared Raw baseline encounters the same 42B tokens multiple times, which likely reinforces the retention of factual world knowledge. In contrast, the L2T setup replaces a portion of these repetitions with structured tasks. For the 1B model, the results indicate that the benefits of linguistic stimulation do not fully compensate for the reduced exposure to raw sequences in knowledge-intensive benchmarks. Thus, to perform well on these benchmarks, larger models need to balance structural induction with factual consolidation.⁶

5 Analysis

Efficacy of Individual Tasks. To isolate the contributions of specific tasks, we pre-train models on 25B task-specific tokens. As shown in Figure

⁶The analysis in Appendix F supports this conclusion by showing that performance on knowledge-intensive benchmarks is sensitive to the ratio of raw text.

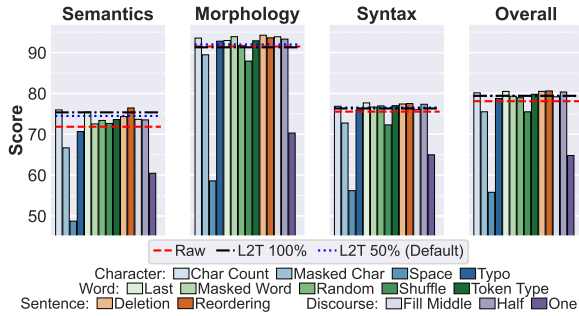


Figure 3: Linguistic competence comparisons on BLiMP between different L2T models trained on specific 25B token single task data.

Data	Numeric Abilities		Conceptual Understanding		Fluid Reasoning	
	Distance	Ratio	Latent Rep.	Zero-Shot	RPM	
500M Share (Disj.)	Raw	0.958	0.826	-0.080	0.172	0.312
	L2T	0.961	0.830	-0.098	0.184	0.366
500M Share (Cont.)	Raw	0.960	0.778	-0.074	0.134	0.292
	L2T	0.958	0.847	-0.087	0.201	0.300
1B Share (Disj.)	Raw	0.947	0.797	-0.128	0.164	0.322
	L2T	0.969	0.870	-0.125	0.201	0.352
1B Share (Cont.)	Raw	0.970	0.833	-0.125	0.137	0.322
	L2T	0.973	0.850	-0.125	0.173	0.298

Table 4: Psychometric evaluation results across cognitive domains. Performance is measured using R^2 for Numeric Abilities, Spearman ρ for Conceptual Understanding, and Accuracy for Fluid Reasoning. **Bold** indicates the superior result for a given scale and data regime.

3, nine tasks (e.g., *Char Count*, *Reordering*) consistently outperform the Raw baseline, suggesting they provide critical structural scaffolding for linguistic acquisition.⁷ In contrast, character-level tasks such as *Space* and *Masked Char* underperform the Raw baseline, with a performance degradation of up to 33 points in Morphology for the *Space* task. We speculate that these tasks might create an unstable training signal when used standalone. Furthermore, while the *One* task largely preserves general performance due to the functional overlap of the objective with standard CLM (Figure 6), it provides insufficient structural signal to elicit linguistic gains. Ultimately, the combined L2T framework consistently surpasses the performance of most single-task models on both BLiMP and general benchmarks. This underscores the robustness of the approach and demonstrates that the integration of diverse tasks elicits complementary strengths, a result that parallels the generalization gains observed when fine-tuning LMs on broad task distributions (Wei et al., 2022; Padmakumar et al., 2022; Sanh et al., 2022).

Generalization beyond Linguistic Competence. We evaluate the impact of L2T on broader cogni-

⁷See Appendix E for the phenomenon-level breakdown.

tive intelligence using the psychometric framework of Shah et al. (2024), covering *Numeric Abilities*, *Conceptual Understanding*, and *Fluid Reasoning*. Table 4 shows that structural stimulation yields consistent gains beyond linguistic competence. For fluid reasoning, L2T substantially enhances Raven Progressive Matrices (RPM) performance (+5.4% for 500M and +3.0% for 1B) in the Disjoint setting. This suggests that L2T enhances the capacity of the model to induce abstract patterns, which is fundamental to fluid intelligence. Furthermore, L2T models consistently demonstrate a stronger Ratio Effect in numeric abilities (e.g., 0.870 for 1B L2T vs. 0.797 for 1B Raw in the Disjoint setting), aligning the representation of magnitude more closely with the processing of humans. Finally, L2T improves conceptual understanding by enhancing zero-shot alignment with the typicality gradients of humans. These results suggest that diverse structured tasks enable the model to develop representations of semantic categorization rather than relying on simple word-level associations.

Discussion. The primary value of the L2T framework resides in the shift of pre-training from the reproduction of surface patterns to structural induction. As demonstrated in Figure 2, L2T models outperform Raw baselines in linguistic competence as early as 5B tokens, establishing the framework as an efficient mechanism for imparting linguistic inductive biases during the initial stages of pre-training. Unlike instruction tuning, L2T induces structure directly from raw text without external supervision. We expect that the field will adopt the L2T paradigm as a foundational objective to ensure that models grasp the underlying linguistic scaffolding. Nonetheless, a direct comparison with supervised augmentation remains a critical direction. Such an analysis would clarify the advantages of the unsupervised approach of L2T relative to methods that rely on human-annotated data.

6 Conclusion

We presented L2T, a pre-training framework that integrates language learning tasks alongside standard next-token prediction. By necessitating the extraction and restructuring of information, these tasks demand processing beyond rote learning, effectively stimulating faster and improved linguistic competence development. Future work will extend L2T to multilingual settings to investigate the learning behavior across languages.

Limitations

Task Scope. The current design of the L2T framework focuses largely on constraints at the level of the sentence. While this effectively targets local syntactic and semantic dependencies, it does not explicitly address broader linguistic phenomena. Future work could expand to include more discourse-level and cross-sentence tasks. Such tasks would allow the model to capture long-range dependencies more effectively.

Model Scale. Due to constraints on computational resources, we restrict our evaluation to models at the 500M and 1B parameter scales. We follow the protocol of Cheng et al. (2024a) by training our models from scratch with a reasonable academic budget of 100B tokens. We acknowledge that the effects of different pre-training objectives might vary at larger scales (e.g., 10B+ parameters), and their investigation is a great opportunity that should be explored by community members that have access to the required compute. We speculate that the structural scaffolding of L2T could be particularly beneficial during the initial stages of pre-training for larger models, as observed in Shah et al. (2024) and Hu et al. (2025). However, the drop in reasoning performance of the Shared 1B model suggests that larger models are likely more sensitive to the size of raw text. Consequently, the application of L2T at scale likely necessitates a careful balance between the volume of raw text and structured tasks to preserve world knowledge, and it may benefit from the use of curriculum learning (e.g., confining L2T data to the initial stage of pre-training) (Benigno et al., 2009; Platanios et al., 2019).

Single Seed. The findings in this study are based on a single training run for each configuration. The high computational cost of pre-training models of 500M and 1B parameters from scratch for 100B tokens made multiple runs prohibitive within the constraints of the available academic infrastructure. However, the consistency of the performance gains across two distinct model scales and two different data regimes (Shared and Disjoint) provides evidence for the robustness of the L2T framework. This indicates that the improvements in linguistic competence result from the systematic impact of structured stimulation rather than sampling variance. Furthermore, the hyperparameters follow the established protocols of Cheng et al. (2024a) to ensure standard training dynamics. To

support verification and the scaling of L2T by the research community, the source code and the data transformation pipeline are publicly available at <https://github.com/gucci-j/l2t>.

Analysis of Individual Tasks. Assessing the efficacy of individual tasks (§5 and Appendix E) relies on experiments conducted at a reduced scale of 25B tokens. Although this volume of data may not provide a comprehensive profile across the entire pre-training duration, it specifically covers the window of maximal development. During this period, LMs typically demonstrate a linear improvement in cognitive and linguistic abilities (Shah et al., 2024). Because performance gains tend to plateau following this initial phase, this interval represents a theoretically grounded window for the assessment of task impact. This targeted analysis enables the identification of essential structural signals while maintaining computational feasibility.

Acknowledgements

We would like to thank Anthony Hughes for the valuable feedback. We acknowledge (1) the use of the University of Oxford Advanced Research Computing (ARC) facility: <http://dx.doi.org/10.5281/zenodo.22558> and (2) the Isambard-AI National AI Research Resource (AIRR) (McIntosh-Smith et al., 2025), which is operated by the University of Bristol and is funded by the UK Government’s Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023]. AY is supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/W524360/1] and the Japan Student Services Organization (JASSO) Student Exchange Support Program (Graduate Scholarship for Degree Seeking Students). MM is supported by the UKRI AI Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. NA is partly supported by the EPSRC [grant number EP/Y009800/1].

References

- Nameera Akhtar. 1999. *Acquiring basic word order: evidence for data-driven learning of syntactic structure*. *Journal of Child Language*, 26(2):339–356.
- Ahmed Alajrami and Nikolaos Aletras. 2022. *How does*

- the pre-training objective affect what large language models learn about linguistic properties? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 131–147, Dublin, Ireland. Association for Computational Linguistics.
- Ahmed Alajrami, Katerina Margatina, and Nikolaos Aletras. 2023. Understanding the role of input token characters in language models: How does information loss affect performance? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9085–9108, Singapore. Association for Computational Linguistics.
- Afra Alishahi. 2011. *Computational Modeling of Human Language Acquisition*, first edition. Synthesis Lectures on Human Language Technologies. Springer Cham.
- Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, and 30 others. 2024. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '24, page 929–947, New York, NY, USA. Association for Computing Machinery.
- Stéphane Aroca-Ouellette and Frank Rudzicz. 2020. On Losses for Modern Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4970–4981, Online. Association for Computational Linguistics.
- Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. 2022. Efficient training of language models to fill in the middle. *arXiv preprint*, arXiv:2207.14255.
- Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, page 41–48, New York, NY, USA. Association for Computing Machinery.
- Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.
- Roger Brown and Jean Berko. 1960. Word association and the acquisition of grammar. *Child Development*, 31(1):1–14.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Susanne Carroll, Merrill Swain, and Yves Roberge. 1992. The role of feedback in adult second language acquisition: Error correction and morphological generalizations. *Applied Psycholinguistics*, 13(2):173–198.
- Tyler A. Chang and Benjamin K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350.
- Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. 2022. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3558–3573, Seattle, United States. Association for Computational Linguistics.
- Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024a. Instruction pre-training: Language models are supervised multitask learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2550, Miami, Florida, USA. Association for Computational Linguistics.
- Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024b. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.
- Jiali Cheng and Hadi Amiri. 2025. Linguistic blind spots of large language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–17, Albuquerque, New Mexico, USA. Association for Computational Linguistics.
- Cheng-Han Chiang and Hung-yi Lee. 2022. On the transferability of pre-trained language models: A study from artificial datasets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10518–10525.

- Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The MIT Press.
- Noam Chomsky. 2002. *Syntactic Structures*. De Gruyter Mouton, Berlin, New York.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try ARC, the AI2 reasoning challenge](#). *arXiv preprint*, arXiv:1803.05457.
- Yiming Cui, Wanxiang Che, Shijin Wang, and Ting Liu. 2022. [LERT: A linguistically-motivated pre-trained language model](#). *arXiv preprint*, arXiv:2211.05344.
- Jennifer Culbertson and David Adger. 2014. [Language learners privilege structured meaning over surface frequency](#). *Proceedings of the National Academy of Sciences*, 111(16):5842–5847.
- Jillian Da Costa and Rui Chaves. 2020. [Assessing the ability of transformer-based neural models to represent structurally unbounded dependencies](#). In *Proceedings of the Society for Computation in Linguistics 2020*, pages 12–21, New York, New York. Association for Computational Linguistics.
- Tri Dao. 2024. [FlashAttention-2: Faster attention with better parallelism and work partitioning](#). In *Proceedings of the Twelfth International Conference on Learning Representations*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Luca Di Liello, Matteo Gabburo, and Alessandro Moschitti. 2022. [Effective pretraining objectives for transformer-based autoencoders](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5533–5547, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yassir El Mesbahi, Atif Mahmud, Abbas Ghaddar, Mehdi Rezagholizadeh, Phillippe Langlais, and Prasanna Parthasarathi. 2023. [On the utility of enhancing BERT syntactic bias with token reordering pretraining](#). In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 165–182, Singapore. Association for Computational Linguistics.
- Jeffrey L. Elman. 1993. [Learning and development in neural networks: the importance of starting small](#). *Cognition*, 48(1):71–99.
- Evelina Fedorenko and Rosemary Varley. 2016. [Language and thought are not the same thing: Evidence from neuroimaging and neurological patients](#). *Annals of the New York Academy of Sciences*, 1369(1):132–153.
- Lukas Galke, Yoav Ram, and Limor Raviv. 2024. [Deep neural networks and humans both benefit from compositional language structure](#). *Nature communications*, 15(10816):1–13.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2023. [A framework for few-shot language model evaluation](#).
- Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. [SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning](#). In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.
- Yuxian Gu, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. [Learning instructions with unlabeled data for zero-shot cross-task generalization](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1617–1634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Qingyan Guo, Rui Wang, Junliang Guo, Xu Tan, Jiang Bian, and Yujiu Yang. 2024. [Mitigating reversal curse in large language models via semantic-aware permutation training](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11453–11464, Bangkok, Thailand. Association for Computational Linguistics.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. [Training compute-optimal large language models](#). *arXiv preprint*, arXiv:2203.15556.
- Jennifer Hu and Roger Levy. 2023. [Prompting is not a substitute for probability measurements in large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.
- Michael Y. Hu, Jackson Petty, Chuan Shi, William Merrill, and Tal Linzen. 2025. [Between circuits and Chomsky: Pre-pretraining on formal languages imparts linguistic biases](#). In *Proceedings of the 63rd*

- Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9691–9709, Vienna, Austria. Association for Computational Linguistics.
- Philip A. Huebner, Elicor Sulem, Fisher Cynthia, and Dan Roth. 2021. [BabyBERTa: Learning more grammar with small-scale child-directed language](#). In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7B](#). *arXiv preprint*, arXiv:2310.06825.
- Gregory D. Keating. 2008. [Task effectiveness and word learning in a second language: The involvement load hypothesis on trial](#). *Language Teaching Research*, 12(3):365–386.
- M. G. Kendall. 1938. [A new measure of rank correlation](#). *Biometrika*, 30(1/2):81–93.
- Payal Khullar, Kushal Majmundar, and Manish Shrivastava. 2020. [NoEl: An annotated corpus for noun ellipsis in English](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 34–43, Marseille, France. European Language Resources Association.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Adhiguna Kuncoro, Chris Dyer, Laura Rimell, Stephen Clark, and Phil Blunsom. 2019. [Scalable syntax-aware language models using knowledge distillation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3472–3484, Florence, Italy. Association for Computational Linguistics.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. [RACE: Large-scale Reading comprehension dataset from examinations](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.
- Tal Levy, Omer Goldman, and Reut Tsarfaty. 2023. [Is probing all you need? indicator tasks as an alternative to probing embedding spaces](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5243–5254, Singapore. Association for Computational Linguistics.
- Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario   aško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, and 13 others. 2021. [Datasets: A community library for natural language processing](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021. [LogiQA: a challenge dataset for machine reading comprehension with logical reasoning](#). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3622–3628.
- Miguel L  pez-Otal, Jorge Gracia, Jordi Bernad, Carlos Bobed, Luc  a Pitarch-Ballesteros, and Emma Angl  s-Herrero. 2025. [Linguistic interpretability of transformer-based language models: a systematic review](#). *arXiv preprint*, arXiv:2504.08001.
- Simon McIntosh-Smith, Sadaf Alam, and Christopher Woods. 2025. [Isambard-AI: a leadership-class super-computer optimised specifically for artificial intelligence](#). In *Proceedings of the Cray User Group, CUG ’24*, page 44–54, New York, NY, USA. Association for Computing Machinery.
- Maggie Mi. 2023. [Mmi01 at the BabyLM challenge: Linguistically motivated curriculum learning for pre-training in low-resource settings](#). In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 269–278, Singapore. Association for Computational Linguistics.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. [Can a suit of armor conduct electricity? a new dataset for open book question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.
- Aaron Mueller, Jason Krone, Salvatore Romeo, Saab Mansour, Elman Mansimov, Yi Zhang, and Dan Roth. 2022. [Label semantic aware pre-training for few-shot text classification](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8318–8334, Dublin, Ireland. Association for Computational Linguistics.
- Florence Myles, Janet Hooper, and Rosamond Mitchell. 1998. [Rote or rule? exploring the role of formulaic language in classroom foreign language learning](#). *Language Learning*, 48(3):323–364.
- Miyu Oba, Akari Haga, Akiyo Fukatsu, and Yohei Osaki. 2023. [BabyLM challenge: Curriculum learn-](#)

- ing based on sentence complexity approximating language acquisition. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 290–297, Singapore. Association for Computational Linguistics.
- Vishakh Padmakumar, Leonard Lausen, Miguel Ballesteros, Sheng Zha, He He, and George Karypis. 2022. Exploring the role of task transferability in large-scale multi-task learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2542–2550, Seattle, United States. Association for Computational Linguistics.
- Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben alal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Amy Perfors, Joshua B. Tenenbaum, and Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338.
- Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.
- Kim Plunkett and Virginia Marchman. 1993. From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition*, 48(1):21–69.
- Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *arXiv preprint*, arXiv:2412.15115.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report.
- Ryokan Ri and Yoshimasa Tsuruoka. 2022. Pretraining with artificial language: Studying transferable knowledge in language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7302–7315, Dublin, Ireland. Association for Computational Linguistics.
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, and 21 others. 2022. Multitask prompted training enables zero-shot task generalization. In *Proceedings of the Tenth International Conference on Learning Representations*.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.
- Raj Sanjay Shah, Khushi Bhardwaj, and Sashank Varma. 2024. Development of cognitive intelligence in pre-trained language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9632–9657, Miami, Florida, USA. Association for Computational Linguistics.
- Elizabeth Spelke. 2022. *What babies know: Core knowledge and composition*. Oxford University Press.
- Zia Tajeddin and Ali Rahimi. 2017. A conversation analysis of ellipsis and substitution in global business english textbooks. *International Journal of Society, Culture & Language*, 5(1):1–14.
- Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes a benchmark to assess the linguistic competence of language models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R.

- Bowman. 2020. [BLiMP: The benchmark of linguistic minimal pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. [Finetuned language models are zero-shot learners](#). In *Proceedings of the Tenth International Conference on Learning Representations*.
- Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. [Crowdsourcing multiple choice science questions](#). In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Zenan Xu, Daya Guo, Duyu Tang, Qinliang Su, Linjun Shou, Ming Gong, Wanjun Zhong, Xiaojun Quan, Daxin Jiang, and Nan Duan. 2021. [Syntax-enhanced pre-trained model](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5412–5422, Online. Association for Computational Linguistics.
- Aditya Yadavalli, Alekhya Yadavalli, and Vera Tobin. 2023. [SLABERT talk pretty one day: Modeling second language acquisition with BERT](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11763–11777, Toronto, Canada. Association for Computational Linguistics.
- Atsuki Yamaguchi, George Chrysostomou, Katerina Margatina, and Nikolaos Aletras. 2021. [Frustratingly simple pretraining alternatives to masked language modeling](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3116–3125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Atsuki Yamaguchi, Hiroaki Ozaki, Terufumi Morishita, Gaku Morio, and Yasuhiro Sogawa. 2023. [How does the task complexity of masked pretraining objectives affect downstream performance?](#) In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10527–10537, Toronto, Canada. Association for Computational Linguistics.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.
- Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. [ReCoRD: Bridging the gap between human and machine commonsense reading comprehension](#). *arXiv preprint*, arXiv:1810.12885.
- Shuai Zhang, Wang Lijie, Xinyan Xiao, and Hua Wu. 2022. [Syntax-guided contrastive learning for pre-trained language model](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2430–2440, Dublin, Ireland. Association for Computational Linguistics.
- Yue Zhang and Stephen Clark. 2015. [Discriminative syntax-based word ordering for text generation](#). *Computational Linguistics*, 41(3):503–538.

Appendix Directory

- **Appendix A: Related Work**
- **Appendix B: Details on L2T: Language Learning Tasks**
 - Character-level
 - Word-level
 - Sentence-level
 - Discourse-level
- **Appendix C: Extended Experimental Setup**
 - Pre-training Data Construction
 - Implementation and Training Details
 - Evaluation Details: BLiMP Benchmark
- **Appendix D: Analysis on Additional General Benchmarks**
- **Appendix E: Efficacy of Individual Tasks**
- **Appendix F: Mixing Ratio of Raw and L2T Data**
- **Appendix G: Qualitative Analysis**
- **Appendix H: License**
- **Appendix I: Use of Generative AI Tools**

A Related Work

A.1 Pre-training and Linguistic Competence

Previous work has questioned the depth of linguistic understanding in LMs (Rogers et al., 2020; Bender et al., 2021; Chang and Bergen, 2024; López-Otal et al., 2025). Empirical analysis reveals that despite proficiency in generating coherent text, models often fail to process distant and complex co-occurrences, such as rhetorical relations (Waldis et al., 2024), and fine-grained linguistic annotation tasks, including noun structures at the phrase level (Cheng and Amiri, 2025). We argue that these shortcomings occur because training signals from standard CLM lack the structural scaffolding necessary for the model to move beyond surface-level statistics.

Consequently, understanding how LMs acquire

linguistic knowledge during pre-training remains a key research area. Evidence indicates that the characteristics of pre-training data play a crucial role. For instance, child-directed language aids grammar induction more effectively than conventional text (Huebner et al., 2021; Yadavalli et al., 2023). Likewise, pre-training on artificial language data designed to model specific structures, such as nesting dependencies, transfers knowledge successfully to natural language tasks (Chiang and Lee, 2022; Ri and Tsuruoka, 2022; Hu et al., 2025). Further, Alajrami and Aletras (2022) suggest that the architecture of the model and pre-training data influence linguistic acquisition more than the objective function. These findings collectively support the data-centric approach of this work to enhance linguistic competence by providing the explicit guidance required to resolve complex linguistic dependencies.

A.2 Enhancing Linguistic Competence of LMs

Prior research explores various strategies to improve the linguistic competence of LMs, typically involving architectural modifications (Xu et al., 2021), auxiliary tasks or objectives (Kuncoro et al., 2019; Xu et al., 2021; Zhang et al., 2022; Mueller et al., 2022; Cui et al., 2022; Guo et al., 2024), curriculum learning (Hu et al., 2025), or data transformation (Guo et al., 2024).

However, research specifically targeting decoder-based LMs remains limited (López-Otal et al., 2025). Existing methods often focus on isolated phenomena or rely on external resources. For instance, Guo et al. (2024) introduce semantic-aware permutation to mitigate the “reversal curse,” but this requires an auxiliary LM and focuses on continual pre-training. Similarly, Hu et al. (2025) utilize formal language data to capture hierarchical dependencies; while this improves generalization, it functions primarily as a warm-up phase using synthetic structures.

In contrast, L2T targets broad linguistic abilities in decoder-based LMs without auxiliary models or external knowledge. By training from scratch, we isolate the impact of our data-centric intervention, demonstrating how structured tasks alone can stimulate the development of linguistic competence during pre-training.

A.3 Self-Supervised Objectives and Data Transformation for Pre-training

Early work explores various self-supervised objectives, particularly for encoder-based models, to improve downstream performance and computational efficiency, and to interpret learned representations (Aroca-Ouellette and Rudzicz, 2020; Yamaguchi et al., 2021; Di Liello et al., 2022; Yamaguchi et al., 2023; Alajrami and Aletras, 2022; El Mesbahi et al., 2023; Alajrami et al., 2023).

More recently, researchers have focused on the transformation of raw text into structured input-output pairs suitable for CLM. These methods include generation of pairs based on predefined self-supervised tasks (Chen et al., 2022), curation of instruction-response pairs using auxiliary models (Cheng et al., 2024a), creation of pseudo-labeled data (Gu et al., 2022), or adaptation of domain text into task formats (Cheng et al., 2024b). These approaches typically aim to enhance general capabilities, improve few-shot learning, or adapt models to specific tasks or domains, often through continual pre-training (Chen et al., 2022).

Our work aligns with the use of data transformation and self-supervision but differs fundamentally in approach and objective. Unlike previous strategies that rely on external models or task-specific datasets for transformation (Cheng et al., 2024a; Gu et al., 2022; Cheng et al., 2024b), our data transformation applies intrinsically to any raw text using predefined rules. Crucially, while prior work often targets improved downstream task performance, our goal is distinct: *to enhance the linguistic competence of LMs during pre-training*.

B Details on L2T: Language Learning Tasks

We introduce 14 language learning tasks (Figure 4) designed to convert raw text into structured input-output pairs. These tasks highlight specific linguistic aspects and necessitate processing capabilities beyond simple next-token prediction. We organize the tasks by linguistic granularity, ranging from local surface dependencies (character/word) to deep syntactic and global coherence structures (sentence/discourse). In the following, we include excerpts from the training samples for each task. [...] signifies that text is omitted for brevity.

B.1 Character-level

B.1.1 Character Count (Char Count)

Char Count counts total characters in the input text. This task promotes fine-grained processing and arithmetic abilities, aiding the handling of morphological variations.

Char Count

Calculate the character count of the following text:

The younger generation can never hope to see again the forest primeval, but the next thing to seeing it with one's own eyes is to see it in Cooper's word pictures, as it was before the ax of the lumberman laid it in ruins. Of Cooper's sea stories, the best is The Pilot, which tells in graphic style of the exploits of John Paul Jones in English waters. [...] To have created Leather-stocking is a passport to enduring fame; yet Cooper added to this typical American backwoodsman Long Tom Coffin, the shrewd Yankee sailor, and a long line of other original characters.

(A) 1833

B.1.2 Masked Character Replacement (Masked Char)

Masked Char reconstructs the original text after replacing a percentage of characters (e.g., 15%) with a mask token. This fosters character-level understanding and morphological processing.

Masked Char

[...] A single g+++ld aiguill+++++te is w++++rn o+++ offic+++r cut unifo++++ms of minor +++etach+++ent+++ . +++ doubl+++ gol+++ aig+++il++++e++++te is wo+++++++ on officer cut un++++forms of major detachment+++ . All +++ersonnel of +++ajor detac+++ments wear +++f++++icer cut uni++++orms.

A: [...] A single silver aiguillette is worn on enlisted cut uniforms of minor detachments. A single gold aiguillette is worn on officer cut uniforms of minor detachments. A double gold aiguillette is worn on officer cut uniforms of major detachments. All personnel of major detachments wear officer cut uniforms.

B.1.3 Whitespace Insertion (Space)

Space restores original spacing to text stripped of whitespace. This enforces word boundary learning and tokenization rules while implicitly modeling grammatical structures defined by spacing.

Space

Femalesmakeof55.5%oftherecordeddeathsand malesmakeup44.5%Whatwastheaveragebirthweightofthenewbornsthat survived?Therewerebirth

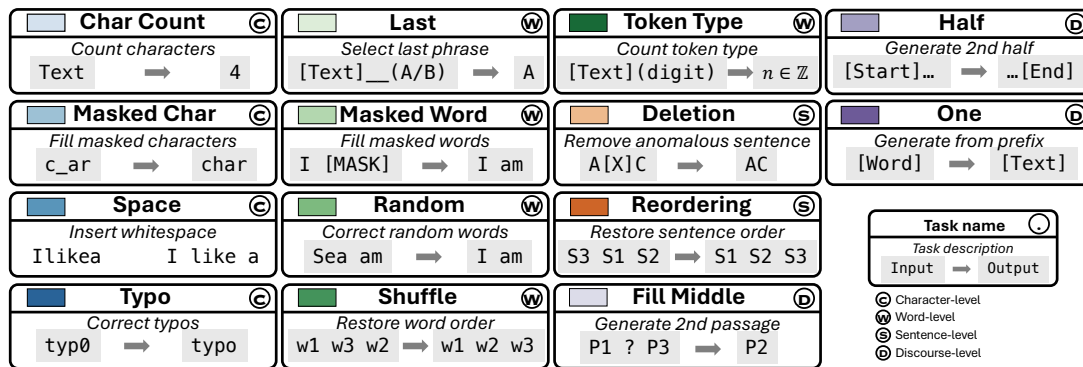


Figure 4: Overview of the 14 language learning tasks. Colors denote linguistic granularity: character (blue), word (green), sentence (orange), and discourse (purple).

weightsrecordedfor nineof the newbornsthat survived.[...] -Theleadingcauseof newbornmortalitywas predationbypumas.Thiswasfollowedby starvation.

Response: Females make of 55.5% of the recorded deaths and males make up 44.5% What was the average birth weight of the newborns that survived? [...] - The leading cause of newborn mortality was predation by pumas. This was followed by starvation.

B.1.4 Typo Correction (Typo)

Typo corrects synthetically introduced typos generated by replacing a random percentage of characters (e.g., 3-8%). This task targets character and word recognition.

Typo

For example, individuals with a first-degree relative (parent, sibling or child) who was diagnosed with lung cancer at an eally age may be at increased risk. [...] "Accessiyility to qhis team df physicians with expertiso in every aspect of diagnosis zo tdeatmlnt wyfl allow fcr more coordinated care, minumibing any scweenfng harms."

For example, individuals with a first-degree relative (parent, sibling or child) who was diagnosed with lung cancer at an early age may be at increased risk. [...] "Accessibility to this team of physicians with expertise in every aspect of diagnosis to treatment will allow for more coordinated care, minimizing any screening harms."

B.2 Word-level

B.2.1 Last Phrase Prediction (Last)

Last selects the correct concluding phrase (the segment following the final stop word) from two candidates. Inspired by Chen et al. (2022), this focuses on the understanding of sentence structure and context.

Last

[...] These savings per person are converted to savings per unit area as follows: On the basis of population estimates (21) from 53 counties in New York State, the median population density was estimated at 103 persons per sq. mi (25th percentile = 67; 75th = 204). Thus, for the areas baited, the savings were calculated at \$156.56 per sq. mi for the first 2 epizootic years (\$1.52 per person x 103 persons per sq. mi), and \$30.90 per sq. mi for the post-epizootic years (\$0.30 per person x 103 persons per sq. mi). Cost-savings data from New Jersey (5) are used in +++?

Options:
and large estates.
the sensitivity analysis.

A. the sensitivity analysis.

B.2.2 Masked Word Replacement (Masked Word)

Masked Word reconstructs text where a percentage of words (e.g., 15%) are replaced with a mask token. Similar to Masked Language Modeling (Devlin et al., 2019), this enhances vocabulary knowledge and contextual inference.

Masked Word

[...] This implies \$\$\$ +++ have a high demand for protein and require optimum temperatures for their metabolisms (()) function at optimal levels necessary for growth ... Energy and protein are ___ by the organism for maintenance, growth and/or reproduction (Staton (()) al. 1986), where energy <> derived from carbohydrates, @@@ and fats. To prevent protein ___ +++ used as an energy source, @@@ energy should be supplied in the diet in [MASK] form of carbohydrates and fat.

A. [...] This implies that they have a high demand for protein and require optimum temperatures for their metabolisms to function at optimal levels necessary for growth ... Energy and protein are required by

the organism for maintenance, growth and/or reproduction (Staton et al. 1986), where energy is derived from carbohydrates, protein and fats. To prevent protein from being used as an energy source, sufficient energy should be supplied in the diet in the form of carbohydrates and fat.

B.2.3 Random Word Replacement (Random)

Random corrects text where 5% to 10% of words are replaced by random vocabulary tokens. Inspired by Yamaguchi et al. (2021), this improves robustness to noise and deepens contextual understanding.

Random

We mortgage find the meaning. [...] If you read locus one theoretical book on the topic of narrative, this one is a testament candidate. While theoretical in perspective, it's written pessimist a manner that will appeal to anyone involved in telling stories.

We cannot find the meaning. [...] If you read only one theoretical book on the topic of narrative, this one is a good candidate. While theoretical in perspective, it's written in a manner that will appeal to anyone involved in telling stories.

B.2.4 Shuffled Word Correction (Shuffle)

Shuffle restores the original order of text with 5% to 10% shuffled words. Also inspired by Yamaguchi et al. (2021), this focuses on syntax understanding and local dependencies.

Shuffle

white loops and buckles permitted the The box to Side worn from either percussion waistbelt or a cartridge box strap. Marines wore a white buff of cartridge box strap that appears to men about 2 1/4 inches wide. [...] Equipped with a Maynard tape primer system, the priming system of the M1855 was similar to the cap guns used by been cross Musket slings were of Black Leather, and officers arms cautioned against "putting muskets in the hands of the be without slings." Officer be Arms There does not seem to be any standard side arm issued to Marine Officers of the Civil War era.

These loops and buckles permitted the cartridge box to be worn from either the waistbelt or a cartridge box strap. Marines wore a white buff leather cartridge box strap that appears to be about 2 1/4 inches wide. [...] Equipped with a Maynard tape primer system, the priming system of the M1855 was similar to the cap guns used by children today. Musket slings were of Black Leather, and officers were cautioned against "putting muskets in the hands of the men without slings." Officer Side Arms There does not seem to be any standard side arm issued to Marine Officers of the Civil War era.

B.2.5 Token Class Count (Token Type)

Token Type counts occurrences of specific word types (e.g., digits, stopwords, punctuation, content words). This encourages classification based on linguistic categories, enhancing the understanding of text composition.

Token Type

Count the punctuation marks in the following passage.

Whereas watching a relaxation video reduces stress, lowers the levels of stress hormones in the blood stream and induces relaxation. [...] Click here to be taken to Comparison Between the Fight/Flight Response and Relaxation Response

A. 36

B.3 Sentence-level

B.3.1 Sentence Deletion (Deletion)

Deletion identifies an anomalous sentence randomly inserted from external context. With equal probability, the model either (a) reconstructs the passage with the anomaly removed, or (b) generates only the anomalous sentence. This fosters the ability to distinguish irrelevant information and maintain discourse coherence.

Deletion

For men, the formal hakama is usually made of heavy black or gray, striped silk. Those designed for women usually come in dark, solid colors, with the exception of the bright red hakama worn by women who tend the Shinto shrines in Japan. [...] Hakama are the traditional garment for priests who reside at Shinto shrines throughout Japan. They are also worn for the practice of archery and for martial arts such as kendo, karate, and aikido. (PDF) STOP, THINK, SPOT FAKE NEWS . Women wear them only for specific occasions and activities, such as graduation ceremonies and traditional Japanese sports. [...] A rustic version of the hakama was also worn by farmers and woodsmen.

A) (PDF) STOP, THINK, SPOT FAKE NEWS .

B.3.2 Sentence Reordering (Reordering)

Reordering restores the original order of a text containing randomly permuted sentences. This encourages understanding of discourse structure, logical flow, and coherence at the sentence level.

Reordering

Assignments (see full descriptions below): 1. All too often, the closer a design is to completion, the more often and drastic the demands for change become. Demands for change may come from the audience, the client, the market, or their own artistic process. Thus, a successful designer learns to rely on an ability to improvise, to respond playfully and creatively, to work with increasingly narrow limitations and assets, while never losing sight of the overall experience goals. [...] In the first semester, the exercises focus on low- and non-technical forms of interactivity, experimenting with games, improvisational theater, fine arts, toys, interactive narrative, and real and virtual play spaces.

A: CTIN 534: Experiments in Interactivity I Instructors: Bernie DeKoven and Tracy Fullerton Course Description: This course is the first in a two-semester sequence which will introduce students to a variety of concepts in interactivity and teach creative techniques for conceptualizing and producing original interactive projects. [...] After the first two classes, which establish the twin focal points of playfulness and improvisation, students will: · experience the fundamentals of improvisation as it applies to both theatrical and interactive development through participation in a variety of exercises · explore fundamental principles of the psychology of fun, the sociology of playfulness and the art of improvisation · address issues of agency, empathy, control, creativity, collaboration and the natural tension that occurs between artist and audience when the audience has a part in making the art Thursday 6:30 – 9:30PM Media Lab, on the second floor of the Evaluation of student performance: a. Assignments (see full descriptions below): 1. In-class design exercises 2.

B.4 Discourse-level

B.4.1 Second Passage Completion (Fill Middle)

Fill Middle generates the missing second passage given the first and third passages of a text. Analogous to fill-in-the-middle (Bavarian et al., 2022), this strengthens the understanding of multi-passage structure and coherence.

Fill Middle

[...] The first confirmed case in Latin America was identified in Brazil on 26 February, although researchers have said there are indications that there were cases there as early as January. Coronavirus has since spread in the world, and has had more than 10,000 confirmed deaths. Chile is reporting thousands of new coronavirus cases each day, with more than 6,000 people dead. [...]

(A) to every country in the region. More than 2.5 million cases have been recorded, and more than 100,000 people have died, according to the European Centre for Disease Control and Prevention. That is fewer

deaths than in the US and Europe, but both cases and deaths may be under-reported. Latin America's two most populous nations, Brazil and Mexico, have seen the highest number of deaths, more than 64,000 and 30,000 respectively. Peru has the sixth highest number of reported cases

B.4.2 Second Half Completion (Half)

Half generates the latter half of a given text. This encourages the learning of discourse coherence, prediction of long-range dependencies, and maintenance of narrative flow.

Half

[...] Your child must fast six hours before an ultrasound to provide a clear view of the organs. During this painless procedure, your child lies quietly while the radiologist moves the ultrasound probe across the abdomen. A cool jelly coats the probe to improve the ultrasound picture quality. Upper Gastrointestinal

A) Endoscopy Exam Compassionate nurses who are experienced in helping children through these tests will start an intravenous line and help administer medications to relax your child and relieve potential pain. During this test, a long, thin telescope – about as wide as a pen – is inserted into the mouth, down the esophagus, and into the stomach and intestine. Through this telescope and its television camera, the gastroenterologist can see the inner lining of the esophagus, stomach and intestine. Small biopsies can be obtained from this lining. Sensors attached to your child also monitor heart rate, blood pressure and blood oxygen levels to ensure safe conditions throughout the procedure. A typical upper gastrointestinal endoscopy takes about 20 minutes. With this test, the physician examines the inside lining of the entire length of the colon or large intestine.

B.4.3 One Word Prefix Generation (One)

One generates the subsequent text given a single word prefix followed by newlines. This specifically trains the initiation and structuring of generation from minimal context.

One

These

Answer: These routes allow visitors to locate works in the Museum, prepare a visit beforehand, further their knowledge of the collection with a thematic route or discover an enjoyable way of introducing the Museum to children through themes such as animals or princesses. Access to the content is through the index of collections or index of artists. There is also the option to save and select works or articles that the user considers most important or relevant in the Favourites section, and to share content on the social networks (Facebook and Twitter) through direct links in the entries on the works. [...]

C Extended Experimental Setup

This section details the construction of the pre-training data, the parameters of the implementation, and the framework used for evaluation.

C.1 Pre-training Data Construction

We construct our pre-training data by combining standard raw text with data generated via our L2T framework.

Data Sampling Strategies. We investigate two configurations for mixing raw text and L2T data. In the Disjoint configuration, we split source documents into two distinct, non-overlapping sets of equal size. One set is used exclusively for standard CLM and the other is transformed into L2T samples. In the Shared configuration, we utilize the exact same source documents for both tasks. This means the pipeline processes every document twice: once as raw text and once to **stimulate** linguistic learning through L2T transformation.

Sample Generation Pipeline. For standard CLM, documents are tokenized and packed continuously. For L2T, source documents undergo the following pipeline:

1. **Segmentation:** Documents are segmented into sentences and grouped into chunks of approximately 512 tokens to ensure samples consist of complete sentences.
2. **Transformation:** One of the 14 tasks (§2) is applied to each chunk. Pairs are formatted as [Input]\n\n[Prefix] [Output] using randomized prefixes for stylistic variation.
3. **Packing and Mixing:** Transformed chunks are concatenated to fill the maximum sequence length and then shuffled with raw text samples. This strategy provides the structural scaffolding necessary to optimize for linguistic competence while retaining world knowledge (Cheng et al., 2024a).

C.2 Implementation and Training Details

Pre-training Data Construction. The hyperparameters for the curation of the L2T data are listed in Table 5. We use **Bling Fire** (v0.1.8) for efficient sentence segmentation. Following Chen et al. (2022), we use varied mask tokens for the Masked Word and Masked Char tasks. The Last task utilizes the stop word list of NLTK (Bird and Loper, 2004) to identify the final segment of the text.

Hyperparameter	Value or Description
Chunk length	512 tokens
Sentence segmentation	Bling Fire (v0.1.8)
Prefix between input and output	{“Answer:”, “Response:”, “A:”, “(A)”, “A)”, “A.”, “”}
Mask token variations	{ “[MASK]”, “_ _ _”, “@@@”, “###”, “+++”, “<<<”, “(())”, “\$\$\$” }
Masking ratio for Masked Word	0.15
Masking ratio for Masked Char	0.15
Replacement ratio for Random	uniform(0.05, 0.1)
Shuffling ratio for Shuffle	uniform(0.05, 0.1)
Typo ratio for Typo	uniform(0.01, 0.08)
False sample for Last & Deletion	Random sampling from previous document

Table 5: Hyperparameters for curating our TL2T pre-training data.

For the Token Type task, each word is classified as “stopword”, “digit”, or “content” via a prioritized procedure: (i) cleaning punctuation and symbols; (ii) matching the lowercase form against the stopword list; and (iii) verifying if the remaining string consists entirely of digits. Any tokens not meeting these criteria are classified as “content”. To count punctuation marks, we utilize the regex pattern: `! "#$%&'()*+,-./:;<=>?@[\\]^_`{|}~`.

To ensure the disambiguation of tasks, specifically to distinguish Token Type from Char Count, a task-specific instruction is inserted randomly at either the beginning or the end of the input, separated by two newline characters (`\n\n`).

Pre-training Details. Table 6 lists the hyperparameters and configuration settings used for pre-training.

Libraries. We preprocess datasets with Hugging Face Datasets (Lhoest et al., 2021, v3.2.0). We use PyTorch (Ansel et al., 2024, v2.3.0), FlashAttention-2 (Dao, 2024, v2.7.3), and Hugging Face Transformers (Wolf et al., 2020, v4.49.0) for pre-training. For evaluation, we use lm-evaluation-harness (Gao et al., 2023, v0.4.8).

C.3 Evaluation Details: BLiMP Benchmark

To measure the linguistic competence of models, we utilize the BLiMP benchmark, which comprises 67 datasets covering 12 linguistic phenomena. Each sample consists of pairs of minimally

Hyperparameters	500M	1B
Hidden size	1024	1728
Intermediate size	4864	4752
Max window layers	24	30
Number of attention heads	24	30
Number of hidden layers	24	32
Number of key value heads	2	4
Rope theta	1,000,000	1,000,000
RMS norm eps	1e-06	1e-06
Attention dropout	0.0	0.0
Tie word embeddings	True	True
Hidden activation	SiLU	SiLU
Initializer range	0.02	0.02
Vocabulary size	32,000	32,000
Tokenizer	Mistral	Mistral
Batch size	256	256
Train steps	200K (for 100B token training), 50K (for 25B token training)	200K
Sequence length	2,048	2,048
Maximum Learning Rate	3e-4	3e-4
Learning rate scheduler	cosine	cosine
Warmup steps	2,000 (for 100B token training), 1,000 (for 25B token training)	2,000
Adam ϵ	1e-8	1e-8
Adam β_1	0.9	0.9
Adam β_2	0.999	0.999
Gradient clipping	1.0	1.0
Weight decay	0.1	0.1
Training precision	BF16	BF16
Computing infrastructure	2 AMD MI300X GPUs (for 100B token training), 2 H100 (96GB) HBM3 GPUs (for 25B token training)	4 H100 (96GB) HBM3 GPUs
Run time for each model	16 days (for 100B token training), 3.5 days (for 25B token training)	12 days

Table 6: Hyperparameters and training costs for each model scale.

different sentences that contrast in grammatical acceptability to isolate specific phenomena in semantics, morphology, or syntax.

Semantics includes two phenomena: quantifiers (Quant), which test restrictions on distribution (e.g., *fewer than* versus *at most*), and negative polarity items licensing (NPI). The latter assesses whether the model possesses knowledge regarding the accurate placement of words such as *any* (e.g., *I did not see anyone.*).

Morphology covers four phenomena: anaphora agreement (Ana Agr), which verifies whether pronouns correctly match the intended referent (e.g., *John saw himself.*); irregular forms (Irregul), which evaluate knowledge of unpredictable word forms (e.g., *go* becomes *went*); determiner-noun agreement (DN Agr), which tests the numerical correspondence between words such as *this* or *these* and nouns; and subject-verb agreement (SV Agr), which confirms that the verb form agrees with the subject (e.g., *he runs* versus *they run*).

Finally, **Syntax** covers six phenomena: argu-

ment structure (Arg Str), which examines the combination of verbs with necessary components (e.g. *eat* must accompany something to be eaten); binding (Bind), , which tests the structural relationship between a pronoun and the antecedent (e.g. *John saw himself* vs. *John saw him*); control/raising (Ctrl Rais), which evaluates syntactic and semantic differences between various predicate types; ellipsis (Ellipsis), which measures whether expressions can be omitted from a sentence; filler-gap (Fill Gap), which assesses dependencies arising from phrasal movement; and island effects (Island), which check constraints on moving sentence elements out of certain grammatical constructions.

D Analysis on Additional General Benchmarks

We conduct an additional evaluation on SuperGLUE (Wang et al., 2019) to analyze the broader impact of the L2T framework on Natural Language Understanding. While linguistic competence on BLiMP consistently improves (§4.1), results on Su-

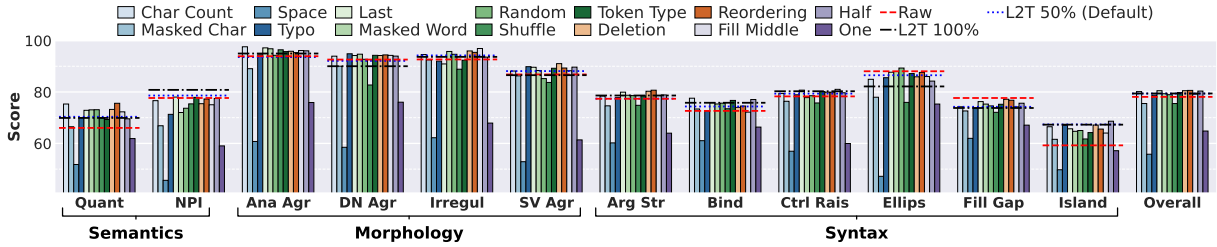


Figure 5: Linguistic competence comparisons on BLiMP between different L2T models trained on specific 25B token single task data.

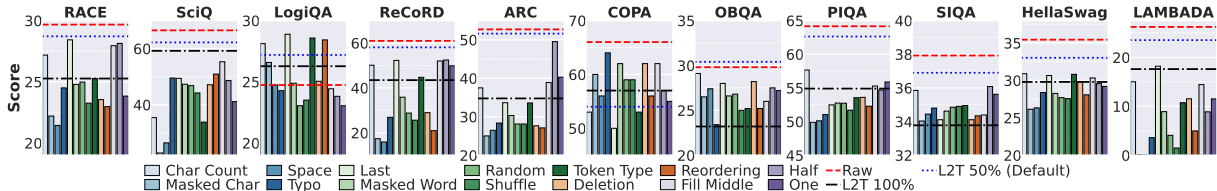


Figure 6: General benchmark performance comparison between different L2T models trained on specific 25B token single task data.

	Data	BoolQ	CB	MultiRC	RTE	WiC	WSC	
500M	Disj.	Raw	53.1	39.7	48.6	56.0	49.7	38.5
		L2T	40.6	31.0	53.3	50.2	47.5	46.2
	Shar.	Raw	56.5	27.3	53.9	48.7	50.0	41.3
		L2T	59.1	31.2	51.1	51.3	47.3	51.9
1B	Disj.	Raw	55.7	21.3	49.4	52.0	50.5	44.2
		L2T	48.4	33.6	51.3	56.0	48.6	45.2
	Shar.	Raw	40.7	33.8	45.0	48.7	50.9	59.6
		L2T	58.0	25.2	49.4	49.5	51.3	42.3

Table 7: Evaluation results on SuperGLUE tasks for models at the 500M and 1B parameter scales. Metrics reported are accuracy (Acc) for all tasks except CB, which uses the F1 score.

perGLUE exhibit variability across different data scenarios. This variability likely originates from the tension between structural induction and factual reinforcement (Fedorenko and Varley, 2016) as discussed in §4.2. Knowledge-intensive tasks, such as BoolQ and CB, depend on the factual repetition inherent in standard causal language modeling. Consequently, performance decreases in the Disjoint scenario because the exposure to raw text is reduced. Conversely, L2T benefits tasks that require structural resolution, such as MultiRC, by enhancing sensitivity to syntactic dependencies beyond simple word probabilities. The size of the model also modulates these effects. The limited capacity of models with 500M or 1B parameters likely results in high sensitivity to the ratio of raw text necessary for general reasoning. Ultimately, these results indicate that L2T serves as a targeted linguistic force multiplier. The framework achieves

higher effectiveness when balanced with sufficient raw text to anchor general factual reasoning.

E Efficacy of Individual Tasks

A closer examination of individual linguistic phenomena (Figure 5) reveals that while almost all tasks (excluding *Space*) outperform the Raw model on Island effects, none succeed on Fill Gap. While the structural scaffolding from the tasks across varying granularity assists in detecting island violations, the complexity of Fill Gap requires targeted signals to capture moved elements and long-distance hierarchical dependencies. Current L2T data likely lacks the specific coverage needed for these complex structures.

F Mixing Ratio of Raw and L2T Data

We investigate the influence of the proportion of raw text relative to L2T data by varying the L2T mixing ratio at 100% (denoting zero raw text), 75%, 50% (the default), and 25% for 500M models using 100B tokens in the Disjoint setting.⁸ Here, a 100% mixing ratio denotes training solely on L2T data.

Linguistic Competence. Results on BLiMP (Figure 8) show that after training, all L2T models perform similarly across linguistic subfields, regardless of the mixing ratio. Differences remain minor, with maximum deltas of 1.1 in semantics, 0.17 in morphology, and 0.73 in syntax. However, early

⁸Due to computational limits, experiments use only the 500M model.

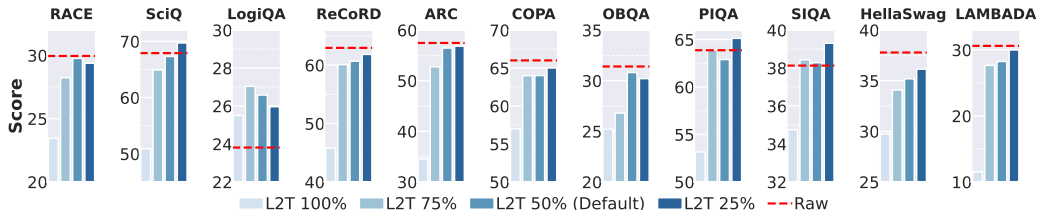


Figure 7: Performance on general benchmarks for 500M models pre-trained with different mixing ratios of standard Raw vs. L2T data for 100B tokens. L2T 100% stands for no standard raw text mixed, i.e. 100% L2T data.

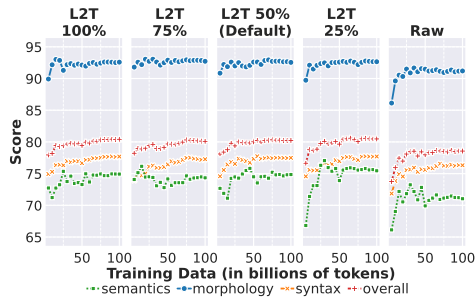


Figure 8: Linguistic competence comparisons by linguistic subfield on BLiMP between Raw and L2T 500M models with different mixing ratios of standard raw text. 100% stands for no standard raw text mixed.

in training (e.g., 5B tokens), performance gains frequently correlate with increased L2T data, except for the L2T 100% setting. For instance, at 5B tokens, L2T 75% outperforms L2T 50% and L2T 25% in morphology (91.8 vs. 90.8 and 89.7) and semantics (74.1 vs. 72.7 and 66.8). This suggests that while the mixing ratio influences the initial learning trajectory, L2T data stimulates linguistic competence regardless of the specific ratio by the end of training.

General Benchmarks. While linguistic gains are stable, evaluation on general benchmarks (Figure 7) underscores the necessity of the raw text proportion (i.e., allocating sufficient training steps to raw text). The L2T 100% configuration, which contains no raw text, exhibits substantial performance drops, such as a 23-point decline on ARC, compared to the Raw model. Increasing the proportion of raw text mitigates this gap; for example, the L2T 75% setting (i.e., containing 25% raw text) differs by only 4.7 points on ARC. Performance generally improves as the proportion of raw text increases (except on LogiQA), a trend supported by Kendall tau correlations (Kendall, 1938) ranging from 0.67 to 1.0. These results demonstrate that while L2T data enhances linguistic competence, raw text remains essential for broad knowledge and reasoning. This aligns with the view of Cheng et al.

(2024a) that mixing raw text is vital for retaining broad world knowledge. Consequently, achieving an appropriate allocation balance between these data types is imperative. Even 25% of L2T data substantially improves linguistic competence (e.g., 1.9 overall gain over Raw in BLiMP), while at least 25% of raw text is essential for robust general capabilities (e.g., an 18-point gain on ARC compared to L2T 100%).

G Qualitative Analysis

We conduct a qualitative analysis to examine the behavior and limitations of models trained on L2T. First, we observe a substantial improvement of 23.2 on the “coordinate structure constraint complex left branch” task within island effects (Island), specifically in the 500M Disjoint setup. This task requires distinguishing between minimally different sentences such as:

Correct: Whose *mice* can Julia bring and Brett notice?
Incorrect: Whose can Julia bring *mice* and Brett notice?

The challenge lies in detecting subtle syntactic violations caused by the misplacement of constituents. We attribute improvements on such tasks largely to the exposure of the model to diverse structural objectives within L2T, which collectively enhance sensitivity to complex syntactic dependencies beyond simple word probabilities.

In contrast, we observe a 7.9 point drop on the “wh vs that with gap long distance” task within filler gap (Fill Gap), which tests whether a long-distance dependency is correctly licensed, as shown below:

Phillip forgot [**Correct:** *what*] [**Incorrect:** *that*] some senator that was escaping from Stacy goes to.

This phenomenon hinges on tracking hierarchical syntactic relationships across intervening clauses. The drop in performance suggests that while L2T effectively captures structural constraints, it continues to struggle with dependencies spanning extensive contexts. This remains an area where even humans perform modestly (75% accuracy)

Category	Dataset	Source Link	License
RC	RACE (Lai et al., 2017)	http://www.cs.cmu.edu/~glai1/data/race/	Custom (Research)
	SciQ (Welbl et al., 2017)	https://allenai.org/data/sciq	CC BY-NC 3.0
	LogiQA (Liu et al., 2021)	https://github.com/lgw863/LogiQA-dataset	No license found
RC+CR	ReCoRD (Zhang et al., 2018)	https://sheng-z.github.io/ReCoRD-explorer/	Apache 2.0 + Internet Archive’s Terms of Use
CR	ARC (Easy) (Clark et al., 2018)	https://allenai.org/data/arc	CC BY-SA 4.0
	COPA (Gordon et al., 2012)	https://people.ict.usc.edu/~gordon/copa.html	No license found
	OpenBookQA (Mihaylov et al., 2018)	https://allenai.org/data/open-book-qa	Apache 2.0
	Social IQa (Sap et al., 2019)	https://huggingface.co/datasets/allenai/social_i_qa	CC BY 4.0
	HellaSwag (Zellers et al., 2019)	https://rowanzellers.com/hellaswag/	MIT
Language Modeling	LAMBADA (Paperno et al., 2016)	https://zenodo.org/records/2630551	CC BY 4.0

Table 8: Summary of Datasets, Sources, and Licenses

and LMs have often faced challenges (Da Costa and Chaves, 2020; Warstadt et al., 2020).

H License

This study uses publicly available datasets with different licenses, as detailed in Table 8. We also use a tokenizer file of Mistral available at [mistralai/Mistral-7B-v0.1](https://huggingface.co/mistralai/Mistral-7B-v0.1), licensed under Apache 2.0. Note that all permit their use for academic research.

I Use of Generative AI Tools

The authors acknowledge the use of LLMs during the preparation of this work. Gemini 3.0 Pro were utilized to find related work and to improve the grammar and clarity of the draft. Additionally, GPT-5 served as a coding assistant for implementation and debugging.